How Many Samples Are Needed

Instructors

Kelly Black
Neptune & Co., Colorado

John Warren
Office of Environmental Information, EPA

2

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
29 MAR 2011		**00-00-2011 to 00-00-2011**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
How Many Samples Are Needed	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Environmental Protection Agency,Office of Environmental Information,1200 Pennsylvania Avenue, N.W.,Washington,DC,20460**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
Presented at the 2011 DoD Environmental Monitoring & Data Quality Workshop (EMDQ 2011), 28 Mar ? 1 Apr, Arlington, VA.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**38**	

Schedule

8:00	Which formula shall I use?
8:30	Choosing the right formula: Estimation
9:15	*Break*
9:45	Practical example: Tasiburn Road
10:30	Choosing the right formula: Decision Making
11:15	Did I take enough samples?
11:30	*Conclude*

3

Purpose of the presentation

- To better understand what formula to use and what assumptions are required
- To gain an understanding of how sample size can change depending on the purpose of the project
- To know if (probably) enough samples have been taken

4

Which Formula Should I use?

Ask the statistician: How many samples?

Me (Statistician): OK, how close do you need the answer to be?
You (Investigator): Maybe within +/- 10ppm

How sure do you need to be?
Pretty sure, is that the 95% thing?

You are talking about the mean aren't you?
I suppose so

What's the variability of the population, i.e., variance?
Don't know

It is a homogeneous population isn't it?
I guess

Can I assume Normality of the population?
Isn't that usual?

What power do you need?
Huh? I'll do it myself!

6

Look in a Textbook

From one standard textbook: $n = \frac{(z_{1-\alpha/2})^2 \sigma^2}{d^2}$

From one EPA document: $n = \frac{(z_{1-\alpha} + z_{1-\beta})^2 \sigma^2}{d^2}$

From another document: $n = \frac{(z_{1-\alpha} + z_{1-\beta})^2 s^2}{d^2} + \frac{1}{2} z_{1-\alpha}^2$

From yet another: $n = \frac{(z_{1-\alpha/2})^2 s^2}{d^2} + \frac{1}{2} z_{1-\alpha/2}^2$

…but what do the terms mean and which one should I use?

7

Explaining the terms used

There are clearly some similarities:

"d" = Within how much does the estimate need to be

"σ^2" = The variance of a population

"s^2" = The variance of a sample

"z" = Relates to the Normal distribution

"1 - α" = How much certainty and related to significance level

"1 - β" = How much certainty and related to statistical power

…but it is still not clear what these terms really mean… 8

What do we know about these terms?

"d" = OK, this we can determine, the "+/-" part of estimation

"σ^2" = The variance of a population (how do we know this?)

"s^2" = The variance of a sample (what sample?)

"z" = Relates to the Normal distribution (i.e., the bell curve)

"1 - α" = Certainty & significance level (is this 95%?)

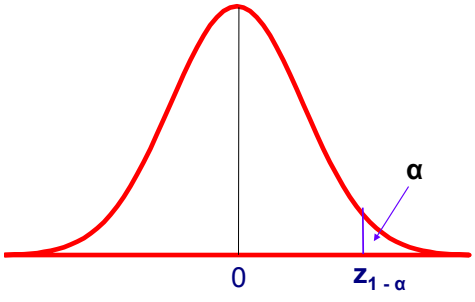
"1 - β" = Certainty & statistical power (not sure about this)

…but it is still not clear where we get these things…

9

Sorting out the terms: $z_{1-\alpha}$

This is the z-value (Normal, bell-shaped curve) that has 1 – α in the "body" of the curve, and α in the "tail"



Level of Significance	(α)	$z_{1-\alpha}$
10%	0.10	1.28
5%	0.05	1.645
2.5%	0.025	1.96
1%	0.01	2.33

10

Sorting out the terms: $z_{1-\beta}$

- This is the same idea as shown with $z_{1-\alpha}$ but is only used when the sample is going to be used for decision making.
- It is related to the statistical power of a specified test (e.g., Student's *t*-test). Power = 1 – β
- It represents the chance of getting it wrong when the Null Hypothesis is false and the Alternative Hypothesis is true.
- You don't need this formula if you are doing estimation.

11

Sorting out the terms: σ^2

This is the variability of the data as defined by the variance (σ^2)

It is very rare that we know this value although we can try and estimate it by learning from similar projects

Sometimes we can get an estimate by making assumptions about the project

The more we can assume, the better the estimate of variability

12

Assumptions and estimating sigma

We can obtain a rough estimate of sigma (from which we then get σ^2) using the maximum and minimum known data values (the Range of the data). Divide the Range by 6 to obtain estimated sigma.

Assumption	Justification	Risk of error
None	Chebychev	0.111
Data > 0	Cantelli	0.100
Unimodal	Vysochanski-Petunin	0.049
Bell-shaped	Normal	0.003

13

Range divided by 6 entails some risk

- If the maximum is really "way off the scale" then the estimate of sigma will be too high and so the number of samples needed will also be high.
- Conversely, if we underestimate sigma we will think we're doing better than we really are as we haven't collected enough samples.

Assumption	Risk	Consequence
None	0.111	Probably sigma underestimated
Data > 0	0.100	Probably sigma underestimated
Unimodal	0.049	Likely a pretty good estimate
Bell-shaped	0.003	Definitely a good estimate

14

The formulae and the assumptions

- **No Assumptions: Chebychev's Inequality**

$$\text{Prob}(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

- **All data > 0: Cantelli's Inequality**

$$\text{Prob}(X - \mu \geq k\sigma) \leq \frac{1}{1 + k^2}$$

- **Unimodal: Vysochanski-Petunin Inequality**

$$\text{Prob}(X - \mu \geq k\sigma) \leq \frac{0.444}{k^2} \quad \text{for } k > 1.633$$

- **Bell-shaped: Normal Theory**

$$\text{Prob}(|X - \mu| \geq 3\sigma) = 0.003$$

15

Now that we know something about the inputs….

From one standard textbook: $n = \frac{(z_{1-\alpha/2})^2 \sigma^2}{d^2}$

From one EPA document: $n = \frac{(z_{1-\alpha} + z_{1-\beta})^2 \sigma^2}{d^2}$

From another document: $n = \frac{(z_{1-\alpha} + z_{1-\beta})^2 s^2}{d^2} + \frac{1}{2} z_{1-\alpha}^2$

From yet another: $n = \frac{(z_{1-\alpha/2})^2 s^2}{d^2} + \frac{1}{2} z_{1-\alpha/2}^2$

… which formula should we use?

16

Choosing the Right Formula (Estimation)

Example: Field sprayed with pesticide

- Ferris Field is almost rectangular, roughly 200 x 150 yds (270,000 square feet, about 6 acres or roughly 20 houses).
- Several years ago it was sprayed with a pesticide now regarded as hazardous. We need to estimate the average level of pesticide present in the top 6″ of soil at Ferris Field.
- Reports from the owner of the site (who wishes to redevelop Ferris Field for residential use) indicate values as high as 90ppb have been recorded. Measurements can be taken accurately down to 0.2ppb.
- The long term exposure level is 50ppb.
- How many samples should we take?

18

Do we have sufficient information?

No

19

Could we make some typical scenarios ?

Yes

Make some reasonable assumptions and see what happens

1. **What are we asked for?** ***"…estimate the average level…"***
   Good, we're talking Mean and no decision-making needed.
2. **Do we know the standard deviation?** ***"…as high as 90ppb…"***
   OK, the maximum seen is 90 and if we assume the pesticide doesn't occur naturally, then 0 must be the minimum.
3. **Within how much should the estimate be?** ***"…exposure level is 50ppb…"*** **OK, let's try +/-10% to start with i.e., +/- 5ppb.**
4. **How sure do they want to be?** ***"… …"***
   They didn't say so let's try 95% to start with.

20

What formula should we choose?

- We should discard all those that demand the population standard deviation (σ) as we don't know it; we have to estimate it from the available data.
- How shall we estimate it? From the maximum and minimum but it depends on assumptions:

Assumption	Justification	Risk of error
None	Chebychev	0.111
Data > 0	Cantelli	0.100
Unimodal	Vysochanski-Petunin	0.049
Bell-shaped	Normal	0.003

Which assumption is most likely to hold?

21

Consider the problem again

- Ferris Field is very large and was sprayed with a pesticide sometime in the past.
- Its unlikely it was pure pesticide but probably mixed with something – could be water, we don't know.
- Fairly sure that it would have been reasonably well mixed although there could be some variations in strength.
- Have no details on how the pesticide was applied so some areas may have been sprayed twice.
- There could be some "hot-spots" but more likely a reasonably uniform contamination with high values, some medium values, and low values.

22

We conclude its probably unimodal

- If it is unimodal (some kind of distribution with a maximum but we don't know the shape) then taking the Range divided by 6 will give us a reasonable estimate of sigma. The Vysochanski-Petunin inequality tells us there's only a 5% chance we're underestimating sigma.
- If we could assume its Bell-shaped (Normal) then there would less than 0.1% chance of underestimating sigma but this would be pushing it!
- Maximum was 90ppb, Minimum has to be 0ppb (actually 0.2ppb as that is the detection level) and so

$$\text{Estimated sigma} = \frac{90 - 0.02}{6} = 14.966$$ i.e., roughly 15

23

Unimodal, applying the formula

- As we estimated sigma from the data it is "s = 15" and so we choose a formula that depends on "s":

$$n = \frac{(z_{1-\alpha/2})^2 s^2}{d^2} + \tfrac{1}{2} z_{1-\alpha/2}^2$$

- The other formula containing "s" doesn't apply when we are doing estimation.
- We can now calculate "n" for various combinations of the input variables α, s, and d.

24

Identifying the pieces

$$n = \frac{(z_{1-\alpha/2})^2 s^2}{d^2} + \tfrac{1}{2} z_{1-\alpha/2}^2$$

- "n": the number of random samples needed
- "d": Within how much (10% of the mean? Which is 5ppm
  20% of the mean? Which is 10ppm)
- "s": The estimated standard deviation (15 in this case)
- "1- α/2": How sure do you want to be (90%, 95%, 97.5%, 99%)

25

What do we have?

For d = 5ppb; here's the number of samples needed :

s	α/2 = 0.10	α/2 = 0.05	α/2 = 0.025	α/2 = 0.01
15	14	26	37	52
18	22	37	52	73
20	27	45	64	90
22	33	54	77	108
25	42	69	98	139
30	60	99	141	199

26

Let's eliminate the unlikely scenarios

The pink scenarios are unlikely to apply to this problem

s	α/2 = 0.10	α/2 = 0.05	α/2 = 0.025	α/2 = 0.01
15	14	26	37	52
18	22	37	52	73
20	27	45	64	90
22	33	54	77	108
25	42	69	98	139
30	60	99	141	199

27

Why "unlikely to apply to this problem"?

- **The α/2 = 0.01 column**
  Choosing such a stringent value is rarely applicable in environmental situations as it is not "life or death".
- **The α/2 = 0.10 column**
  Choosing such a relaxed value is rarely done except for controlled exploratory studies. α/2 means that the chance we miss the true mean is 20% (α = 0.20) which is high.
- **The s = 25 through 30 rows**
  For "s" to be 25 would imply the maximum was 6 x 25 = 150, which is nearly half as much again than the maximum we observed. As this is a long way past the maximum we conclude it is unlikely to be this high.

28

What's left?

- The sample size needed is between 26 and 77 depending on our estimate of "s" and choice of "$\alpha/2$", assuming that the allowance in final estimate (d) is 5.
- For "d = 10" (previously "d" = 5) we recalculate and find:

s	α/2 = 0.05	α/2 = 0.025
15	8	11
18	11	15
20	13	18
22	15	21

- If we assume it is unlikely the budget for the pesticide investigation will allow for more than 30 samples, the conclusion is:

29

What to report

If you want the estimate to be +/- 10% (i.e., 5ppb)

Take 30 samples:

- You'd be pretty sure the estimate is good (roughly 95% confidence)
- If the assumptions were off just a little then you couldn't be sure as you haven't enough samples.

Take 20 samples:

- Even if the assumptions were you couldn't be sure as you haven't taken enough samples.

If you want the estimate to be +/- 20% (i.e., 10ppb)

Take 30 samples:

- Very sure the estimate is good (exceeds 95% confidence) even if the assumptions were not true. Probably should make "d" smaller otherwise more than enough samples have been taken.

Take 20 samples:

- Pretty sure the estimate is good even if the assumptions are not true (95% confidence)

30

Applying to Ferris Field

Ferris Field was 200 x 150 yards, equal to 270,000 sq ft.

If we take 30 random samples, each sample would have to represent 9000 sq ft equal to about 32 x 30 yards. Is it feasible that a single sample could be representative of such a large area? (13,500 sq ft if 20 samples are taken)

The formula assumes a random sample could we locate random locations at Ferris Field? Have access to GPS?

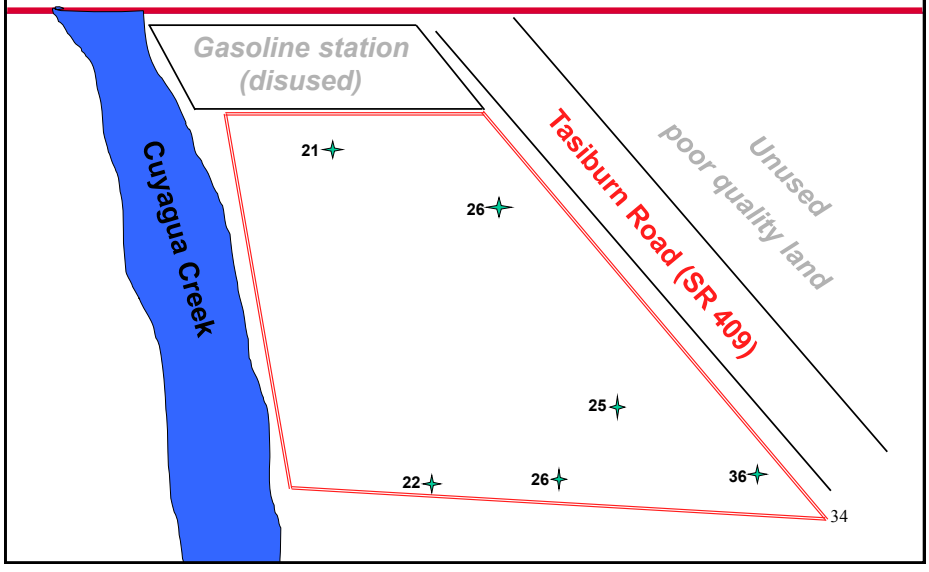
Can we cut down on the number of samples sent for analysis? Could we use field sampling techniques? Can we change sampling schemes? Are there alternatives?

31

32

Example: Tasiburn Road

Class Example: Tasiburn Road Development Project



Tasiburn Road Development Project

- Tasiburn road is a 2-lane State-maintained road in western Arkabama. Prior to being made SR 409 it was a dirt road with easy access both sides.
- About 20 years ago waste oil was sprayed on dirt roads as a dust suppressor – the oil contained *Glomerator*.
- *Glomerator* is now recognized as a contaminant that is a cause for concern.
- The project developer must determine the level of *Glomerator* at the site before taking subsequent action.

35

Tasiburn Road data and information

- The Tasiburn Road Development is bounded by Cuyagua Creek on the west, a disused gasoline station to the north, SR 409 to the east and a fence to the south.
- Traffic density is fairly light and automobile pollution low.
- There may be some leakage from underground storage tanks remaining at the disused gasoline station.
- Cuyagua Creek is shallow, average depth 12" but subject to spring flooding on an irregular basis
- Core samples for 6 locations gave *Glomerator* readings: 21ppm, 22ppm, 25ppm, 26ppm(2), and 36ppm (mean = 26ppm, variance = 28.4ppm^2, standard deviation = 5.329ppm).

What the developer must achieve

- The developer must convince the State Regulator that the estimated mean level of *Glomerator* remaining at the project is accurate to within 2 ppm.
- Also present evidence that the samples results have credibility and will be accepted by the State Regulator.
- Show that a reasonable (defensible) argument may be made that sufficient samples to characterize the project have been taken.
- Show that the estimated mean meets commonly held standards of certainty.

37

How many samples does the developer need?

From one standard textbook: $n = \frac{(Z_{1-\alpha/2})^2 \sigma^2}{d^2}$

From one EPA document: $n = \frac{(z_{1-\alpha} + z_{1-\beta})^2 \sigma^2}{d^2}$

From another document: $n = \frac{(z_{1-\alpha} + z_{1-\beta})^2 s^2}{d^2} + \frac{1}{2} z_{1-\alpha}^2$

From yet another: $n = \frac{(z_{1-\alpha/2})^2 s^2}{d^2} + \frac{1}{2}(z_{\alpha/2})^2$

Make a recommendation and suggest the number of samples needed.

38

Choosing the Right Formula (Decision Making)

Example: Field sprayed with pesticide

- Ferris Field is almost rectangular, roughly 200 x 150 yds (270,000 square feet, about 6 acres or roughly 20 houses).
- Several years ago it was sprayed with a pesticide now regarded as hazardous. We need to determine if the average level of pesticide present in the top 6″ of soil at Ferris Field exceeds the long term exposure level of 50ppb.
- We need to take sufficient samples such that if the true level of pesticide is 60ppd or more, we would be 90% sure we could find this.
- How many samples should we take?

40

Almost the same problem as previously encountered but important differences

Previously:
"We need to estimate the average level…"

Now:
"We need to determine if the average level of pesticide… exceeds the long term exposure level of 50ppb"

Now:
"We need to take sufficient samples such that if the true level of pesticide is 60ppb or more, we would be 90% sure we could determine this"

41

We will use the same assumptions

- **Previously, by using the assumption the data was unimodal, we reached the conclusion that an estimate of sigma was 15.**
- **We then used a formula to find "n" (we concluded that it was 20 – 30 depending on which one the budget would allow).**
- **For decision making ("determine if") with criteria for how certain we need to be ("would be 90% sure"), we should choose an appropriate formula.**
  - **It will be similar, but more complex, than the one used for estimation.**

42

Unimodal, applying the formula

- As we estimated sigma from the data it is "s = 15" and so we choose a formula that depends on "s":

$$n = \frac{(z_{1-\alpha} + z_{1-\beta})^2 s^2}{d^2} + \tfrac{1}{2} z_{1-\alpha}^2$$

- But what is $z_{1-\beta}$?
- It is related to the statistical power of the test when the Alternative Hypothesis is true
- Need to talk about Null and Alternative Hypotheses

43

Null and Alternative Hypotheses

- The Null Hypothesis (H_o) represents the "baseline" condition. We hold on to the Null until faced with overwhelming evidence (data) that shows it can't be true, in which case we choose the Alternative Hypothesis (H_A).
- In this specific case:
  - H_o: The average (mean) level ≤ 50ppm
  - H_A: The average (mean) level > 50ppm
- Note that we do not specify what exactly it is under the Alternative, just that it exceeds what the Null says it is.

44

Null, Alternative, α, and β

- The error rates, α and β, are defined as:
- "α" is the probability of rejecting the Null Hypothesis when it is really true.
- "β" is the probability of accepting the Null Hypothesis when it is really false.
- In this specific case:
  "α" is the chance you say the mean level is above 50ppm when really it is below 50ppm.

  "β" is the chance you say the mean level is below 50ppm when really its above 50ppm.

45

Null, Alternative, α, and β

- "α" is decided at the very start, often set at 10% or 5%.
- "β" is specified for certain values for H_A and then the number of samples needed to achieve this calculated.
- For our example, "α" was not specified, so we chose 5%.
- For our example, "β" was specified, but indirectly:
  "if the true level was 60ppm or more, we would like to be 90% sure we could detect this".

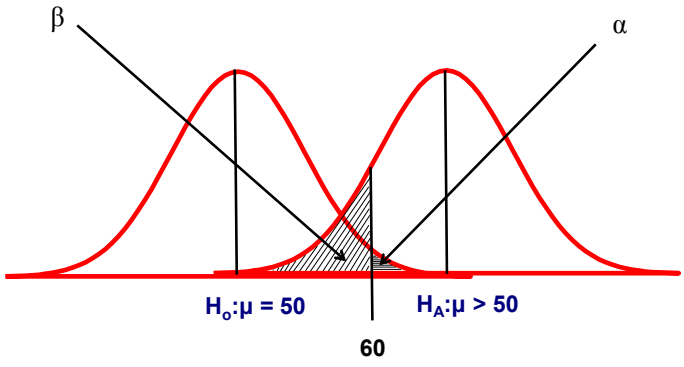
  90% sure we're right equals 10% chance that we're wrong.

  In math terms:

  $1 - \beta = 0.90$, therefore $\beta = 0.10$

46

How $z_{1-\alpha}$ and $z_{1-\beta}$ are related



$1-\alpha$ is the area under H_o up to 60

$1-\beta$ is the area under H_A from 60

47

Identifying the pieces

$$n = \frac{(z_{1-\alpha} + z_{1-\beta})^2 s^2}{d^2} + \tfrac{1}{2} z_{1-\alpha}^2$$

- "n": the number of random samples needed
- "d": Within how much (difference between Ho = 50ppm, and the specified H_A = 60ppm)
- "s": The estimated standard deviation (15 in this case)
- "1- α": How sure do you want to be (90% and also 95%)
- "1- β": Related to statistical power (90%, $z_{1-\beta}$= 1.28 for 90%)

48

What do we have?

For d = 10ppb; here's the number of samples needed :

s	α = 0.10	α = 0.05
15	16	21
18	23	30
20	28	36
22	33	43

This is for β = 10%

49

Comparison of number of samples: estimation versus decision making

For d = 10ppb:

Estimation

s	α/2 = 0.10	α/2 = 0.05
15	8	11
18	11	15
20	13	18
22	15	21

Decision making

s	α = 0.10	α = 0.05
15	16	21
18	23	30
20	28	36
22	33	43

This is for β = 10%

Note that decision making needs more samples

50

What to report

If you took 30 samples:

- You'd be quite sure (over 95% confidence) that enough samples have been collected unless the assumptions are well off in which case you'd be unsure as not enough samples have been taken.

If you took 20 samples:

- You'd be pretty sure (95% confidence) that enough samples have been taken if the assumptions are true. If the assumptions are not true then you'd be unsure as not enough samples have been taken.

51

Applying to Ferris Field

Ferris Field was 200 x 150 yards, equal to 270,000 sq ft.

If we take 30 random samples, each sample would have to represent 9000 sq ft equal to about 32 x 30 yards. Is it feasible that a single sample could be representative of such a large area? (13,500 sq ft if 20 samples are taken)

The formula assumes a random sample could we locate random locations at Ferris Field? Access to GPS?

Can we cut down on the number of samples sent for analysis? Could we use field sampling techniques? Can we change sampling schemes? Are there alternatives?

52

Did I Take Enough Samples?

You don't know

- You calculated the number of samples needed, took the required samples (including those for Quality Assurance controls), but are you sure you took enough?
- You don't (and may never) know.
- Could you use the results from the sample to answer this?
- The answer is "Yes, maybe, if you're lucky…"

54

"Back-estimating" number of samples

- We take the sample standard deviation and use the formula to calculate "backwards" to get "n". Let's use the example for decision making:

$$n = \frac{(z_{1-\alpha} + z_{1-\beta})^2 s^2}{d^2} + \tfrac{1}{2} z_{1-\alpha}^2$$

- We agreed on 30 samples (before QA additional samples) based on a sigma of 15 (actually we noted it could have been as high as 18). Suppose we did this and the standard deviation of the sample was 19.35.
- Using the formula "backwards" yields "n" = 34

55

What to conclude

- Well, using the sample standard deviation we should have taken 34 instead of 30. Probably the 30 we took is enough, a difference of 4 is not important.
- But, it can be shown that a confidence interval for the unknown standard deviation is approximately between 15.41 and 26.01.
- Using 15.41 in the formula gives "n" = 22
- Using 26.01 in the formula gives "n" = 60
- So, not too much help; best advice is to calculate, document, back-estimate, and use common-sense.

56

Reducing Sample Size by using Stratification

57

Sampling & Field Variability

Total variability of any project has two parts:
Field Variability + Measurement variability

(Between unit) (Within unit)
Field variability greatly exceeds Measurement variability

We can reduce the number of samples by reducing Field variability. Stratification makes more homogeneous areas within the entire problem and we can find the required number of samples for each area and then combine for an overall total.

58

Simple Random Sampling

- **Used when population variability is high**
- **Simple in concept and provides proper data (theoretical support) for statistical data analysis**
- **Is the basic building block of more complicated (but more effective) probability-based designs**
- **Difficult to find individual sample locations**
- **Often demands a large number of samples**

59

Stratified Sampling

- **The target population is divided into contiguous sub-populations (strata) of approximately the same variability**
- **Sampling locations are selected within each strata using some sampling design**
- **Needs information on what criteria are meaningful in defining the strata boundaries**
- **Greatly reduces variability**

60

Grid (Systematic) Sampling

- **Collecting samples according to a specified pattern at regular intervals**
- **Can yield more precise estimates of population parameters than other sampling designs**
- **Easy to explain and implement and provides uniform coverage of site or project**
- **Can be biased if the sampling grid pattern or the regular frequency of taking samples coincides with any pattern of contamination**

61

Example: Littlewood Site

- **Site is suspected to be contaminated with arsenic through its production work**
- The site was a multi-purpose factory
- Preliminary information indicate a range in values 0.2ppm to 33ppm

62

Project objective

- The project objective is to estimate the mean arsenic level (of the site and the individual areas) with minimal uncertainty.
- Uncertainty can be measured by precision or variability in the conclusion.
- The number of samples needed will be proportional to the variability encountered, high variability demanding more samples than low variability
- We are just interested in the field aspect – not the measurement as this is the responsibility of others

63

The size of the project

- The map is drawn to scale and so it may be deduced that the entire site is about 6000 sq yds
- For clean-up purposes the site will be divided into operational units size 1500 sq yds (roughly the size of a football field divided by 4)
- Therefore there are 4 operational units (total) in this site to be investigated through sampling

64

Preliminary observations

A very rough estimate of the total variability can be deduced from the range in data values:

$$\text{estimated } s = \frac{\text{range}}{6}$$

$$= \frac{33.0 - 0.2}{6}$$

$$= 5.5$$

The estimated total variance (s^2) is then $(5.5)^2$ which is 30.25 and approximately equal to 30

65

The sample size per operational unit

$$n = \frac{z^2_{1-\alpha/2}\, s^2}{d^2} + \tfrac{1}{2} z^2_{1-\alpha/2}$$

$z_{1-\alpha/2}$ = how sure we need to be (95%)

d = how close we need to be (+/- 2ppm)

n = the formula gives 21.6

i.e., We need 22 per operational unit together with around 4 QC samples to estimate other components

66

The cost for the project

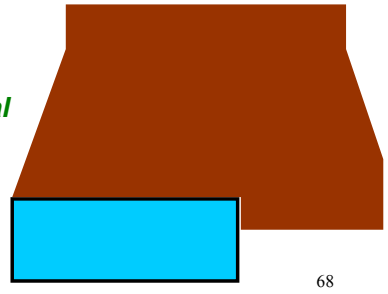
- Total samples = 4 x (22 + 4) =104
- Square grid across the site
- Cost to analyze each sample $1,000
- Total cost is then $104,000
- But the budget is only $35,000

More work is needed as clearly we can only afford a maximum of 35 samples

67

Can reduce variability through stratifying

- Records show that the site contains two distinct areas. A rectangular area where production was conducted and a much larger area containing administration offices and a parking lot.
- **Contamination in the large area believed to be relatively low**
- ***You are improving the conceptual model of operations at the site***



68

Impact of stratification on sample size

- With reduced variability, fewer samples are needed to achieve the same precision
- The high value (33ppm) most likely came from the Production Area so the variability estimate is probably valid in that region
- Production Area (1430 sq yds) is approximately the same as an operational unit (1500 sq yds)
- This would require 26 samples from the Production Area (recall the formula does not consider target population size)

69

Using the same parameters as before

- Due to budget constraints this would leave only 9 samples to be taken from the Admin/Parking area
- However, does the Admin/Parking area have the same variability as the Production area? Probably not…
- Suppose the maximum expected in Admin/Parking was 10ppm, using the formula would demand only 4.3 (5 plus QC samples) and so we would be within budget.

70

Conclusion for Littlewood

Need 22 + 4 = 26 samples for Production Area

Need 5 + 1 = 6 samples for Parking/Admin

As the budget allowed for 35 total, the remaining 3 could be allocated to the Production Area, thus improving the quality of the resulting estimates.

71

Overall Conclusion

- Formulae found in books assume simple random sampling
- Stratification can reduce costs and improve precision
- More complex sampling designs can produce better estimates than simple designs but need more planning

72

How Many Samples Are Needed

Conclusions

- The number of samples needed is dependent on:
  - The purpose of the project
  - d ➔ how close the estimate needs to be
  - σ^2 ➔ the variance of the population of interest
  - s^2 ➔ the variance of a sample
  - z ➔ expected distribution
  - α ➔ significance level and certainty
  - β ➔ statistical power and certainty
- It is important to confirm after sampling that you have a sufficient number of samples to meet your objectives.
- The number of samples necessary can be greatly reduced with clever sampling designs.

74

Want more information?

- ***Guidance on Choosing a Sampling Design for Environmental Data Collection QA/G-5S***
  (www.epa.gov/quality)
- ***Statistical Methods for Environmental Pollution Monitoring***
  Richard O. Gilbert
- ***Sampling***
  Steven K. Thompson

75

Still need more help?

Kelly Black
Neptune & Company
kblack@neptuneinc.org

John Warren
Office of Environmental Information, EPA
warren.john@epa.gov

76